Bottom-Up Multiple Row Addition Algorithms for the Biclustering-Problem

Hyung-Won Koh  
ISAS - Institute for Analytical Sciences  
44139 Dortmund, Germany  
Email: koh@isas.de

Lars Hildebrand  
University of Dortmund  
44221 Dortmund, Germany  
Email: lars.hildebrand@udo.edu

Abstract—The discovery of locally and significantly correlated subpatterns within a two-dimensional dataset has recently become quite popular and is amongst others addressed by methods solving the biclustering problem. The preservation of a particularly defined degree of homogeneity between elements within a bicluster plays a key role in the search procedure. A prominent quantity is known as the mean squared residue. Most approaches use such measures only to evaluate the found solutions a posteriori, instead of incorporating them directly into the search procedure. This work proposes a pairwise distance function related to the mean squared residue and includes this measure into two new multiple enrichment algorithms. The impact is demonstrated empirically by the enrichment of bicluster sets of a popular method and by the enrichment of randomly chosen single rows.

I. INTRODUCTION

In terms of grouping genes into correlating subsets, we distinguish between clustering and biclustering approaches. Clustering in general tries to partition the set of genes regarding the expression values of all conditions, as done for example by hierarchical clustering and k-means clustering approaches [1]. Assuming that genes may participate in many different biological processes simultaneously only for a particular set of conditions, conventional clustering approaches may not be sensitive and/or specific enough to find and present correlations between genes in an appropriate and comprehensive manner.

In recent years, a lot of different biclustering algorithms and methods addressing this drawback have been introduced. They concentrate on uncovering sub-patterns within the matrix preserving a certain degree of correlation on one or both dimensions of the data set. A survey of several well known biclustering approaches has been published by [2]. In general, they distinguish between five different classes of biclustering approaches:

1) Iterative Row and Column Clustering Combination  
2) Divide and Conquer  
3) Greedy Iterative Search  
4) Exhaustive Bicluster Enumeration  
5) Distribution Parameter Identification

The work of this paper is focused on the third of the listed classes: Greedy Iterative Search. Briefly described, biclustering methods within this class traverse the search space of solutions by adding or removing rows or columns in a greedy manner. The word greedy means that such algorithms are based on locally optimal, greedy decisions. Therefore, not all possible solutions are considered, resulting in essential time savings but often to the cost of suboptimal results.

In the context of gene expression analysis, many of these approaches within the class Greedy Iterative Search are based on a predefined similarity score which allows to quantify the degree of homogeneity within a bicluster. One of the earliest measurements in this context is known as the mean squared residue [3], taking into account the single expression values of the elements, the mean values of its rows, the mean values of its columns and the mean value of the whole bicluster itself. Many approaches which are based on this homogeneity function follow a rather top-down than a bottom-up strategy. Some of them integrate greedy manipulation strategies [4], others traverse the search space in a probabilistic manner [5]. Only a minority has been recognized as strategies following a bottom-up approach at all [6]. Thus all of the mentioned approaches use the homogeneity measurement only as an additional tool to quantify their results a posteriori. By introducing a pairwise homogeneity distance function, this work tries to establish bottom-up enrichment algorithms integrating quality measurements directly into the enrichment process in order to improve the accuracy of a biclustering method in general.

II. RELATED WORK

Cheng and Church were one of the first who introduced the term "biclustering" in the context of expression data analysis [3]. They also introduced the mean squared residue as a homogeneity measurement and proposed heuristic algorithms exploiting mathematical properties of the mean squared residue. Since then several different biclustering approaches have been proposed, for example random walk strategies [7], evolutionary algorithms [5], [6] and parameter distribution identification [8], [9].

The biclustering problem is taken to be NP-complete [2], so strategies solving this task should aim only at approximating the optimal solution in order to save time and space resources. Theoretical aspects of the mean squared residue have only slightly or not even at all been analyzed. Many approaches incorporate or extend the heuristic algorithm of Cheng and Church with slightly or no changes. Cho et al. figured out theoretical aspects of the mean squared residue for the columns and for the rows explicitly, and applied k-means clustering on
Finally, a bicluster and equally, the mean value of column \( j \)

**General case of a finite two-dimensional matrix**

In the following we will consider the with the genes and the columns refer to the conditions of a floating-point number. The rows are then in general associated

Given a bicluster \((I, J)\) the base of a row, the base of a column and the base of a mean and the mean of the whole bicluster are also known as

**Definition 1: bicluster**

Given a \( n \times m \) data matrix \( A \) with the set of rows \( N \) and the set of columns \( M \), we denote the element of row \( i \) and column \( j \) of the matrix as \( a_{i,j} \). The row mean, or the mean value of row \( i \) is then given as

\[ a_{i,J} = \frac{1}{|J|} \sum_{j \in J} a_{i,j} , \]

and equally, the mean value of column \( j \) is denoted as

\[ a_{I,j} = \frac{1}{|I|} \sum_{i \in I} a_{i,j} . \]

\( a_{I,j} \) refers to the mean value of the whole bicluster given as

\[ a_{I,J} = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} a_{i,j} . \]

Finally, a bicluster \((I, J)\) is defined as a subset of rows \( I \subseteq N \) and a subset of columns \( J \subseteq M \). The row mean, the column mean and the mean of the whole bicluster are also known as the base of a row, the base of a column and the base of a bicluster respectively [5].

**Definition 2: mean row and mean column of a bicluster**

Given a bicluster \((I, J)\), the mean row \( m_r \) of \((I, J)\) is defined as

\[ m_r(I, J) = \{ a_{I,0}, a_{I,1}, \ldots, a_{I,|J|-1} \} . \]

\( m_r \) stands for the \(|J|-\)dimensional vector maintaining the mean values of all columns, and is therefore clearly defined for each bicluster. Note that the mean row \( m_r \) differs from the row mean, the mean value \( a_{i,j} \) of row \( i \). Same considerations hold for the mean column of a bicluster.

In order to quantify the homogeneity within the elements of a bicluster, the **mean squared residue** \( H(I, J) \) has been introduced by [3] as:

**Definition 3: mean squared residue**

Given a bicluster \((I, J)\), the mean squared residue \( H(I, J) \) is defined as

\[ H(I, J) = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} r(a_{i,j})^2 \]

with

\[ r(a_{i,j}) = a_{i,j} + a_{I,j} - a_{i,J} - a_{i,j} \]

representing the residue of element \( a_{i,j} \). Additionally the row residue \( d_r(i) \) of a row \( i \in I \) of a bicluster \((I, J)\) is defined as:

\[ d_r(i) = \frac{1}{|J|} \sum_{j \in J} r(a_{i,j})^2 . \]

The definition of the column residue \( d_c(j) \) of a column \( j \in J \) is analogously given as

\[ d_c(j) = \frac{1}{|I|} \sum_{i \in I} r(a_{i,j})^2 . \]

The mean squared residue measures the coherence of all elements within a bicluster, with low values indicating a high correlation. A bicluster \((I, J)\) is called a perfect bicluster if \( H(I, J) = 0 \), and a bicluster with \( H(I, J) \leq \delta \) is called \( \delta \)-bicluster. \( \delta \) describes the tolerated level of deviance within a bicluster motivated by technological boundaries of precision and by the occurrence of systematic and non-systematic bias in experimental measurements in general. For this reason, most approaches search for \( \delta \)-bicluster rather than for perfect ones.

Besides the degree of homogeneity, a bicluster \((I, J)\) can also be characterized by measuring its information content. One often used function is known as the **mean row variance** \( \text{Var}_r(I, J) \) [2], defined as:

**Definition 4: mean row variance**

Given a bicluster \((I, J)\), the mean row variance \( \text{Var}_r(I, J) \) is defined as

\[ \text{Var}_r(I, J) = \frac{1}{|I|} \sum_{i \in I} v_r(i) \]

with

\[ v_r(i) = \frac{1}{|J|} \sum_{j \in J} (a_{i,j} - a_{i,J})^2 \]

as the variance of row \( i \).

The **mean column variance** is defined analogously.

**Definition 5: residual distance**

Given the set of rows \( N \) and the set of columns \( M \) of an
$n \times m$ input matrix $A$, the residual distance $d_r(i, k)$ between any pair of rows $(i, k) \in I \times I$ is defined as:

$$d_r(i, k) = \frac{1}{|J|} \sum_{j \in J} r(a_{i,j}, a_{k,j})^2$$  \hspace{1cm} (9)

with

$$r(a_{i,j}, a_{k,j}) = r(a_{k,j}) - r(a_{i,j}) = a_{k,j} - a_{k,j} + a_{i,j} - a_{i,j}$$ \hspace{1cm} (10)

and $r(a_{i,j}, a_{k,j})$ known as the residue of element $a_{i,j}$ and $a_{k,j}$ respectively (see equation (4)). The residual distance between any pair of columns $d_c$ can be defined adequately as

$$r(a_{i,j}, a_{i,l}) = r(a_{i,l}) - r(a_{i,j}) = a_{i,l} - a_{i,l} + a_{i,j} - a_{i,j}.$$  \hspace{1cm} (11)

The residual distance function $d_r$ measures the distance between two rows or two columns, whereby the outcome is related to the homogeneity function mean squared residue. The question of what kind of relation exists and how it can be exploited to establish biclustering algorithms will be answered in detail in the following sections. Note that the residual distance defined here equals the non-negative real-valued function $d$ which occurs as $d(b_i, b_k)$ in Lemma 1 and Lemma 2 of [3].

III. METHODS

First of all, we can observe that the Node Addition algorithm of [3] is a combination of two single procedures: a row addition part and a column addition part. We will consider these parts separately and focus on the row addition part in detail. In [3] only rows $i$ fulfilling $d_r(i) \leq H(I, J)$ are added to the bicluster $(I, J)$. In other words, the row residue $d_r(i)$ of row $i$ has to be less than or equal to the current mean squared residue $H(I, J)$ in order to add $i$ to the bicluster. As a result, the mean squared residue of a bicluster will stay constant or decreases after executing the Node Addition algorithm. It concludes that this algorithm is impracticable for any bottom-up approaches searching for a bicluster with $\delta > 0$, since $H(I, J) = 0$ holds for every bicluster maintaining just one row or one column (without proof). The following theorem leads to a proper solution of the aforementioned problem:

**Theorem 1:** Let $A$ be an $n \times m$ matrix of elements in $\mathbb{R}$, let $N$ be the set of rows and $M$ be the set of columns of $A$, and let $I \subseteq N$ be a subset of rows and $J \subseteq M$ a subset of columns of $A$. Then it holds that

$$\frac{1}{|I|} \sum_{i \in I} d_r(i, k) \geq \frac{1}{|I|} \sum_{i \in I} d_r(i)$$

**Proof:** If we consider the space $S \subseteq \mathbb{R}$ and define the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ as $f(x, y) = \frac{1}{|S|} \sum_{x \in S} (y - x)^2$, we observe that $f$ minimizes if $y$ equals the mean of $S$, formulated as

$$\min_{x \in S, y \in \mathbb{R}} f(x, y) = f(x, \frac{1}{|S|} \sum_{x \in S} x).$$  \hspace{1cm} (12)

If we note that $d_r(i, k) = d_r(k, i)$, the proof succeeds as follows:

$$\frac{1}{|I|} \sum_{i \in I} d_r(i, k) \geq \frac{1}{|I|} \sum_{i \in I} d_r(i)$$

\[
\Rightarrow \frac{1}{|I|} \sum_{i \in I} \frac{1}{|J|} \sum_{j \in J} ((a_{i,j} - a_{i,J}) - (a_{k,j} - a_{k,J}))^2
\]

$$= \frac{1}{|I|} \sum_{i \in I} \frac{1}{|J|} \sum_{j \in J} ((a_{i,j} - a_{i,J}) - (a_{k,j} - a_{k,J}))^2$$

$$\geq \frac{1}{|I|} \sum_{i \in I} \frac{1}{|J|} \sum_{j \in J} ((a_{i,j} - a_{i,J}) - (a_{i,j} - a_{i,J}))^2$$

$$= \frac{1}{|I|} \sum_{i \in I} d_r(i)$$

**Corollary 1:** In consequence to theorem 1 given any bi-cluster $(I, J)$, any row $i \in I$ and any $|J|$-dimensional vector $k \in \mathbb{R}^{|J|}$, it also holds that

$$H(I, J) = \frac{1}{|I||J|} \sum_{i \in I} \sum_{j \in J} r(a_{i,j})^2 \leq \frac{1}{|I|} \sum_{i \in I} d_r(i, k)$$

The basic meaning of theorem 1 and corollary 1 is that we now can estimate the homogeneity of a bicluster based on $d_r(i, k)$. Additionally, we can calculate $d_r(i, k)$ for any row $i$ without calculating the mean values of the columns. As a result, we can formulate a first multiple row addition algorithm as:

**Algorithm 3.1: Integrated Multiple Row Enrichment (IMRE)**

**Input:** a real-valued matrix $A = (N \times M)$, a $\delta$-bicluster $(I', J)$, homogeneity bounds $\delta_1$ and $\delta_2$ with $\delta_1 \leq \delta \leq \delta_2$, positive integer values $K_1, K_2 \in \mathbb{N}$.

**Output:** a $\delta$-bicluster $(I, J)$ with $|I| \geq |I'|$.

**Algorithm:**

1. Compute $m_r(I', J)$, $a_{I', J}$, and $d_r(i, m_r(I', J))$ for all rows $i \in N$, and let $I''$ be an empty set of rows.
2. Add all rows $i' \in N, i \not\in I'$ to the bicluster $(I', J)$, if $d_r(i', m_r(I', J)) \leq \delta_1$ holds.
3. Iterate step 1 and step 2 until no further improvement can be observed or the maximal number of $K_1$ iterations is reached.
4. Compute $m_r(I', J)$, $a_{I', J}$ and $d_r(i, m_r(I', J))$ for all rows $i \in N$.  
5. Sort the set of rows  

$$I'' = \{i'' | i'' \not\in I' \land d_r(i'', m_r(I', J)) \leq \delta_2\}$$

in ascending order of $d_r(i'', m_r(I', J))$.
6. Add in ascending order all rows $i'' \not\in I'$ to $I'$, for which

$$\frac{1}{|I'| + 1} \sum_{i \in I'} d_r(i, m_r(I', J)) + d_r(i'', m_r(I', J)) \leq \delta$$
We can formulate an algorithm including $s_i$ as follows:

**Algorithm 3.2: Weighted and Integrated Multiple Row Enrichment (WIMRE)**

Input: A real-valued matrix $A = (N \times M)$, a bicluster $(I, J)$, a score function $s_i$, a positive real value $\beta$, a homogeneity bound $\delta$, a set of rows $I$ and a set of columns $J$.

Output: A score function $s_I$.

1. Computed the mean variance $\text{Var}(v, r)$ of the initial bicluster $(I, J)$ for all rows $v \in I$ and columns $r \in J$.

2. Add all rows $v \in I$ if $\text{Var}(v, r) < \delta$.

3. Sort the set of rows in ascending order of $s_i$ for all rows $v \in I$.

4. Add in ascending order of $s_i$ all rows $v \in I$ to the bicluster $(I, J)$ if $\text{Var}(v, r) < \beta$.

5. Iterate step 2 to 6 until no further improvement can be observed or the maximal number of $K$ iterations is reached.

The time complexity of algorithm 3.2 is in $O(K \cdot n \cdot m)$, if an enrichment by $\beta$ is not only based on the residual distance, but on the row variance as well. This is done by introducing a score function $s_I = \text{Ic}(I, J)$, where $\text{Ic}(I, J)$ is the current score function for enriching the bicluster $(I, J)$. By defining the score function $s_I$, we can formulate an algorithm including $s_i$ as follows:

**Algorithm 3.3: Integrated Multiple Row Enrichment (IMRE)**

Input: A real-valued matrix $A = (N \times M)$, a set of columns $J$, a score function $s_i$, a homogeneity bound $\delta$, a set of rows $I$, and a set of columns $J$.

Output: A score function $s_I$.

1. Computed the mean variance $\text{Var}(v, r)$ of the initial bicluster $(I, J)$ for all columns $r \in J$.

2. Add all columns $r \in J$ if $\text{Var}(v, r) < \delta$.

3. Sort the set of columns in ascending order of $s_i$ for all columns $r \in J$.

4. Add in ascending order of $s_i$ all columns $r \in J$ to the bicluster $(I, J)$ if $\text{Var}(v, r) < \beta$.

5. Iterate step 2 to 6 until no further improvement can be observed or the maximal number of $K$ iterations is reached.

The time complexity of algorithm 3.3 is in $O(K \cdot n \cdot m)$. The correctness of this runtime concludes due to the consideration that $\text{Ic}(I, J)$, the correctness of the variance bounds discussed in this algorithm, can be calculated in time $O(n \cdot m)$ and that sorting objects needs time $O(n \cdot m)$.
TABLE I

Properties of the three bicluster shown in figure 1. The listed values are rounded to the third decimal place.

<table>
<thead>
<tr>
<th>#</th>
<th>original</th>
<th>IMRE</th>
<th>WIMRE</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>85</td>
<td>97</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td>41</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>0.08</td>
<td>0.08</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.106</td>
<td>0.105</td>
<td>0.107</td>
</tr>
<tr>
<td>17</td>
<td>64</td>
<td>232</td>
<td>183</td>
</tr>
<tr>
<td></td>
<td>32</td>
<td>32</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>0.088</td>
<td>0.088</td>
<td>0.088</td>
</tr>
<tr>
<td></td>
<td>0.345</td>
<td>0.26</td>
<td>0.358</td>
</tr>
<tr>
<td>76</td>
<td>46</td>
<td>440</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>0.069</td>
<td>0.069</td>
<td>0.069</td>
</tr>
<tr>
<td></td>
<td>0.148</td>
<td>0.099</td>
<td>0.184</td>
</tr>
</tbody>
</table>

the values as indicated above. The same strategy has been followed for the second dataset.

Figure 1 shows three arbitrarily chosen bicluster from the CC bicluster set on yeast before and after applying IMRE and WIMRE. We can observe that both succeed in enriching the bicluster. In detail, both algorithms ended up in slightly improvements concerning the homogeneity, but only WIMRE was able to improve the mean row variance of the bicluster, whereas IMRE led to a lowering of it. Table I provides the corresponding numerical results. Additionally, both bicluster sets mentioned above have been enriched totally by the proposed methods on both data sets. The results for CC are shown in figures 2 and 3 and for RandomSet in figures 4 and 5 each for both data sets. As intended, the average number of rows after enrichment is clearly higher for IMRE in all cases (see the top of figures 2 – 5). However, the high enrichment of rows also comes to the price of a decreased mean row variance (see bottom of figures 2 – 5), which might not be surprising since the variance is not considered during the enrichment of IMRE at all. WIMRE compromises between these values of interest in dependence to parameter γ. As demanded, both algorithms maintain the homogeneity of each bicluster for the CC bicluster set, as can be observed in the middle chart of figure 2. The homogeneity of the bicluster from the ath dataset even improves (see middle of figure 3). If we have a more detailed look on the enrichment performance of WIMRE with parameter γ = 0.75, we can observe that the improvement in the homogeneity is accompanied by an increase of sometimes more than twice the amount of rows while simultaneously almost keeping the mean row variance. This emphasizes the decisive superiority in both precision and thoroughness of the proposed approaches on the given datasets. The enrichment performance of RandomSet further shows that all proposed algorithms are indeed able to perform a bottom-up enrichment from the very beginning of just a single row (see figures

![Graphs and images](image-url)

Fig. 1. Enrichment of the 6th (top row), 17th (center row) and of the 76th (bottom row) bicluster of the Cheng and Church algorithm (left column) by IMRE (center column) and by WIMRE (right column). The gray lines indicate the expression values of the genes belonging to the original bicluster, the black lines indicate enriched genes. Table I provides the corresponding numerical results.
Fig. 2. Enrichment results for all 100 bicluster from the Cheng and Church bicluster set on the yeast matrix in descending order. Shown are the number of rows (top), the mean squared residue (middle) and the mean row variance (bottom) for each bicluster.

Fig. 3. Enrichment results for all 32 bicluster from the Cheng and Church bicluster set on the ath matrix in descending order. Shown are the number of rows (top), the mean squared residue (middle) and the mean row variance (bottom) for each bicluster.

V. CONCLUSIONS AND OUTLOOK

Considering the impact of the proposed algorithms IMRE and WIMRE, we may conclude that both are quite promising enrichment methods with regard to the mean squared residue. By the proposed algorithms we are able to boost size, homogeneity and also information content of each bicluster.
measured as the row variance. Due to space limitations, evidence is given empirically only by enriching the bicluster set of the Cheng and Church algorithm and randomly chosen initial bicluster for two datasets. Continuing tests on other bicluster sets have already been performed and could confirm the conclusions made.

In general, the applicability of the proposed methods is not restricted to an afterwards optimization of results gained by other approaches. Rather, they have the potential to become parts in more sophisticated frameworks as precise and thorough enrichment procedures, for example as local search methods within an evolutionary approach to perform biclustering on their own. Compared to several well known approaches, preliminary tests of such frameworks resulted in essential im-
provements regarding the simultaneous optimization of overall coverage, average size and average variance. The analysis and establishment of such frameworks will be the authors subject in the future.

ACKNOWLEDGEMENT

The authors thank Frank Klinker and Thomas Hofmeister for fruitful discussions and suggestions about the mean squared residue. This work has partially been funded by the LEIBNIZ-GEMEINSCHAFT within the excellence initiative “pact for research and innovation: Integrated Cell Analysis”.

REFERENCES


